Least Squares Fit to an Arbitrary Function Using Excel

A Data Analysis Technique

Tara Falcone

Introduction

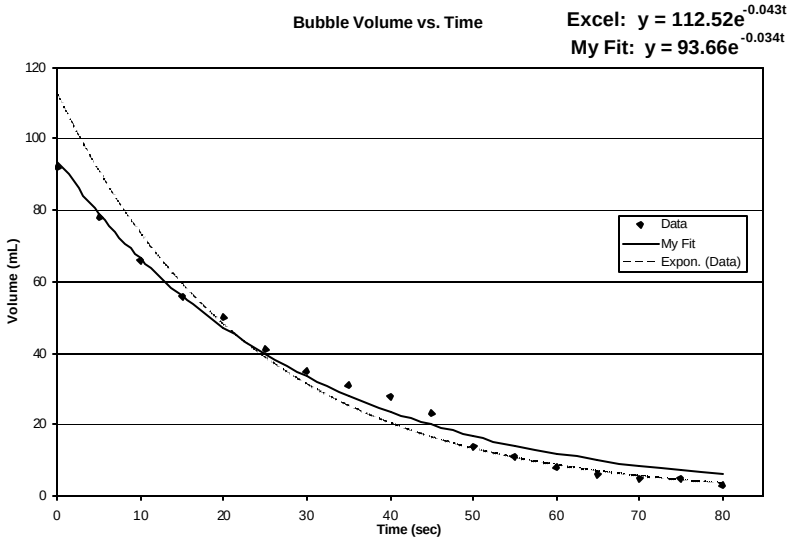
A major concept that deserves more emphasis in science classrooms is data analysis. Physics is an experimental science based on scientific method. It is a science searching for cause and effect relationships in nature. When studying a specific system, one first hypothesizes a theory to relate the chosen variables to the behavior of the system. One tests this theory by performing an experiment and collecting data. This data is analyzed in order to see how well the experimenter's original theory describes the observed behavior of the system. Data analysis is the key to determining whether or not the experimenter should have faith in their current theory, or alter their theory to better fit nature.

A spreadsheet program, like Excel, can be used in order to perform a relatively simple, yet good, analysis of data. It is widely available in schools and homes, and is therefore a good tool for teaching data analysis in high school physics classes. This knowledge will be incredibly valuable to high school students as they take it with them into their experimental science classes in college and hopefully, these concepts will be used throughout their future scientific careers.

With the click of a few buttons, Excel can automatically perform a fit to data. Hence, it is valid to question why one should learn how to analyze the data for themselves. Well, one can answer their own question by comparing their fit to Excel's automatic fit.

Warning: Do not trust Excel's *Add Trendline* command! Because it is not known how it works to calculate a fit, one cannot trust its results!

A good example of this warning is seen in the following depiction:



This data was taken from the exponential decay of soda bubbles experiment. It is clearly seen how Excel's trendline fit and the fit I did to the same exact data varies greatly from one another. The fit I produced is a better description of the behavior of the soda bubbles over time than Excel's exponential fit.

Set-up

In order to begin analyzing data, one obviously must enter the data into Excel. This can be done by typing in the data manually, or by reading it in from a saved text file. The later option is done by first opening a worksheet and selecting the cell in which you want to put the data from the text file. Then, under the *Data* menu, go to *Import External Data*, and then click on *Import Data.* In the *Files of Type* box, click *Text Files*, and in the *Look In* list, locate and double-click the text file you want to import as an external data range. Follow the instructions in the *Text Import Wizard* to specify how you want to divide the text into columns, and then click *Finish*. One can return the data to the *Existing Worksheet* or in a *New Worksheet* in your workbook. In the same dialogue box, formatting and layout options can be set for the imported data by clicking on *Properties*.

When entering data into Excel, always leave some rows free at the top of the worksheet (about 7 or 8) and immediately label your data columns appropriately and with the right units. This will make things a lot easier.

Three columns of data should be entered: one for the independent variable (x coordinate), one for the dependent variable (y coordinate), and one for the error in the dependent variable. At the very top of the worksheet (where empty space was intentionally left), the fitting function should be typed as a convenient reminder. Below that, label individual cells with appropriate parameter names. Underneath each label, you may want to enter an arbitrary number, such as 1, for setup purposes of the next few steps. These cells are where the actual values of the parameters will appear.

Next, create a column next to your data for the fitted function. This is done by selecting the first cell in the column (beneath the label cell) and typing an equal sign followed by the mathematical expression of the function, referring to the proper cell addresses of the function variables and for the values of the parameters (i.e. A3, B12). A cell address can easily be obtained by clicking on the cell you want to refer to. To correctly enter the values of the fitting parameters, each element of the cell address must be preceded by a dollar sign (i.e. $A$3, $B$12). The dollar signs are needed in the parameter addresses so that Excel knows to always refer to those cells as parameters, while the variable cell values change. Excel regards address referral as relative to the selected cell (two left, one down), unless the cell address is preceded by a dollar sign, in which case the referral is absolute (the cell: A3).

For mathematical expressions such as sine, cosine, exponential, etc. use the proper Excel command, which can be found under *Function* in the *Insert* menu, or in the *Help* menu. After correctly entering the formula into the first cell, press return to obtain the calculated value. Then, select the cell again by clicking on it, and position the mouse over the square in the lower right corner where the cursor turns into a cross. Click again and hold

the mouse button while pulling the box down until you reach the number of cells corresponding to the number of data points.

The next column to set up is the column of normalized residuals, which is the deviation of the fit value from the experimental value divided by the uncertainty in the experimental value $\{(y\text{-}fit(y))/error(y)\}$. Enter this mathematical formula using the correct cell addresses and pull down to enter the formula for each data point, similar to what was done previously.

The definition of χ^2 is the sum of the squares of the normalized residuals. Up at the top of the spreadsheet near the parameter cells, label a cell for χ^2. In the cell underneath the label type an equal sign followed by the command **SUMSQ(***argument***)**, where the argument in the parentheses correspond to the values of the normalized residuals. An easy way to refer to the addresses in the argument is to click and highlight the whole column of values of normalized residuals. Notice how when this is done Excel only enters the addresses of the top cell and the bottom cell of the selected column separated by a colon (i.e. A1:A10) instead of identifying each cell address (i.e. A1, A2, …, A10). Cell addresses separated by a comma are distinct, while those separated by a colon are continuous.

Next, label the cell adjacent to χ^2 for the normalized χ^2, which is χ^2 divided by the number of degrees of freedom. The number of degrees of freedom is defined by the number of data points minus the number of fitting parameters. Enter the correct corresponding formula under the normalized χ^2 label to obtain the value. The closer the normalized χ^2 is to the value of 1, the better the fit describes the behavior of the data.

Now, you should have the fitting formula, the parameters, the χ^2, normalized χ^2, and five columns completely set up to start performing the fit with Excel.

Fitting

To find the best fit to describe the behavior of our data we must minimize χ^2. This is done using the *Solver* routine in Excel. *Solver* may or may not already be installed on the computer. If it appears under the *Tools* menu it is installed. If it does not, go to *Tools*, then *Add-Ins*, and check the box for *Solver Add-In* and click *ok*. The installation CD may or may not be necessary for this.

The *Solver* routine is set to perform a hybrid grid-gradient search for a minimum in the χ^2 hyper-surface in parameter space. To begin the fit, an educated guess for the values of the parameters must first be made. The initial values for the parameters are important because the *Solver* search routine could get stuck in a local minimum of the χ^2 surface, while looking for the absolute minimum.

Once guesses have been made for the parameters, go to *Solver* under the *Tools* menu and the *Solver Parameters* submenu will appear. *Set the Target Cell* as the value of the χ^2 cell which we want to minimize (min). This is easily done by clicking on the box with the red arrow in it next to the entry box, and then clicking on the necessary cell. Minimization of χ^2 is achieved by changing the values of our parameters. Thus, in the *By*

Changing Cells entry, enter the cell addresses containing the values of the parameters in a similar way. There is no need to change the default settings; obtain the result by simply clicking *Solve.* Several tries with the Solver routine while manually adjusting the parameters, if necessary, ought to give the absolute minimum of the χ^2 surface. If necessary or desired, one can also subject the search to constraints. In this way, Excel can be set to perform a search where specified parameters are equal to, greater than or equal to, or less than or equal to the values specified. In the *Subject to the Constraints* box, click on *Add* and define constraints by selecting the cell of the specific parameter and setting the desired limits.

Once minimizing χ^2, look at the corresponding value of the normalized χ^2. Again, the closer this value is to 1, the better the fit describes the data. Also, look at the corresponding values of the normalized residuals. Ideally, the distribution of these values should be completely random. If a distinguishing pattern among the normalized residuals exists, and/or if the χ^2 is much greater or less than 1, then one should consider another kind of fitting function or a combination of functions.

Error in the Parameters

Once the values of the parameters have been obtained by fitting, the uncertainty in the parameters can then be found. This is done by performing a χ^2 variation for each parameter independently of the others. In order to investigate the uncertainty in a particular parameter, one must do several solver routines where a value of χ^2 is found for small variations on the value of the fitted parameter. By obtaining a set of values of χ^2, each corresponding to a specific parameter value around the fit value, one quickly discovers a parabolic relationship between the value of the parameter and χ^2. This relationship is parabolic because one is indeed searching around a minimum (Excel finds the best fit by searching for values of the parameters where χ^2 takes on an absolute minimum value- the method of least squares). Using this information, the error in the parameter is obtained by finding the values of the parameter where the corresponding χ^2 is exactly one more than the value of the minimum χ^2. These exact values can be found by first fitting a parabola to the χ^2 vs. parameter plot, and then solving for the intersection points of this equation with the line of value exactly equal to one plus the minimum χ^2 value. Solving for exact numerical values of these roots may be done using a mathematics program, such as Maple or Mathematica, but sometimes one can estimate these values well by looking at the graphical representation of the behavior. The absolute value of the difference between the parameter values at these points and the minimum parameter are equal to each other, as well as to the uncertainty in the fitted parameter value.

To find the uncertainty in each one of the parameters, one must do a separate analysis for each parameter. These analyses involve manipulating the original worksheet created to find the best fit parameter values. To begin this process, the first thing that must be done is to go back to fitting worksheet, run the solver through again using the discovered best fit values for the parameters, but this time, when the *Solver Results* box appears, click on *Save Scenario* and name this scenario something like “best” or “original”. This allows one to save these best fit values for the parameters and allows for the ability to load these values back into the worksheet at a future time.

Next, one must decide an interval to use in adjusting the parameter value in order to investigate around the minimum. Four intervals above and below the minimum, along with the minimum value should yield a nice investigation of nine points. When doing this, keep in mind that the aim of this exercise is to search for the values of the parameters yielding a χ^2 which is exactly greater than the minimum χ^2 by one. Therefore, the largest χ^2 values obtained should not be too much greater than the minimum value.

Once these intervals have been figured out, each value must be placed one at a time into the original fitting worksheet where the solver must be run in order to find the χ^2 corresponding to each parameter value. When running these solver routines, set the *Target Cell* to the χ^2 cell which one wants to minimize like previously done. The main difference in these solver routines is that in the *By Changing Cells* box, one must enter the addresses of all the parameters **except** for the cell address containing the value of parameter you are investigating (which is one of the steps about the minimum value). Clicking on *Solve* will return new values of the other parameters and of the χ^2. The new value of the χ^2 is what should be noted. This χ^2 value can be copied and entered into the worksheet being used to find the error in the parameter. This process needs to be repeated for each one of the eight steps around the minimum.

Now that a table containing variations of the minimized parameter and their corresponding χ^2 values have been obtained, a parabolic curve can be fit to these nine points. This is done in the same fashion as the original fit worksheet made previously. This new setup should contain four columns, one for the variance on the parameter, one for their corresponding χ^2 values, one for the parabolic fit referencing the correct cell addresses, and one for the deviation between the corresponding χ^2 values and the fit values. The parabolic fit should be of the form $y = c(x - p)^2 + (\chi^2)$, where p is the value of the parameter obtained from the original fit, (which one is trying to find the error on) χ^2 is the corresponding minimized value obtained from the original fit using the specific p value of the parameter, and c is a fitting parameter which one is now trying to find. Enter this equation referencing the appropriate cell addresses (remember to use dollar signs for the addresses containing the values of c, p and (χ^2)) and pull down this first box to automatically fill the formulas for each data point. In the next adjacent column, enter the formula for the difference between the corresponding χ^2's and the fit (corresp. χ^2 – fit), and fill in for each point.

Next to the cell for the parameter c, make a cell (and label) for the χ^2 of this parabolic fit. The formula that should be entered in this cell is the sum of the squares of the values in the deviation column (**SUMSQ(***deviation***)**). After this is set up, make a first guess for the value of c and then run the solver routine where you are again minimizing the χ^2 of the fit by changing the value of the parameter c. Once the value of c is obtained, the formula of the parabolic curve is known, and the error in the parameter can be found by solving this equation for the values of the parameters that yield a χ^2 of one greater than the minimum χ^2. The deviation of these values of the parameter from the minimum value of the parameter gives the uncertainty. This whole procedure should be repeated separately for every parameter that an uncertainty needs to be found in.

Graphical Presentation

Graphical presentation of data is very convenient because a picture is in fact worth a thousand words. Excel allows one to make customized data charts relatively easily. This is done by clicking on the *Chart Wizard* icon in the shortcut button menu (the one that looks like a bar graph with blue, yellow and red columns), choosing *XY Scatter*, double clicking on the picture of a point plot, going to the *Series* tab, hitting *Add*, and then selecting the independent and dependent variables by clicking on the box with the arrow beside each entry and then highlighting the columns of values in the worksheet corresponding to the appropriate variables. Multiple graphs can be put on the same chart by adding multiple series of data to the particular chart in this way. This is convenient when showing how well the fit curve conforms to the data points. Following through the *Chart Wizard* by hitting *Next* allows one to label the title and the axes of the chart, and allows the option of displaying or not displaying a legend. At the end of the dialogue box, one can select to embed the chart into the worksheet currently being worked in, or to place the chart in a sheet of its own.

Many options can be manipulated by simply clicking on a data point in a series and then right clicking and selecting *Format Data Series* (or by double clicking on a data point). Under the *Patterns* tab, one can change the color of the points and have an option of adding a smooth line connecting the points, in a selected color. One can choose to represent individual data sets as points on a curve, as points alone, or as just a curve. The size and shape of the point markers can also be varied, along with the thickness and pattern of the curve. Under the *X Error Bars* and *Y Error Bars* tabs, one can add error bars to points in the vertical and horizontal directions if desired. Error bars in the vertical direction are most commonly represented on data analysis graphs. To add these, go to the *Y Error Bars* tab, select the box labeled *Both* under *Display*, check off custom, use the arrow boxes to select and highlight the column of cells corresponding to the values of the uncertainty in the dependent variable in both the + and – entries, and then hit *Ok*. There are also additional options in the other tabs which can be figured out by trial and error. Each data series can be customized independently by selecting a point in the specific data series and then right clicking to obtain its unique formatting options (or by double clicking on a point in the data series). One can also double click on the X and Y axes themselves, in order to obtain formatting options for each. With Excel, very nice customized charts can be made. What is even nicer about Excel is that one can learn how to make such charts by double clicking on what one desires to adjust, and figuring out what each of the various options that appears does by tinkering with the options themselves.

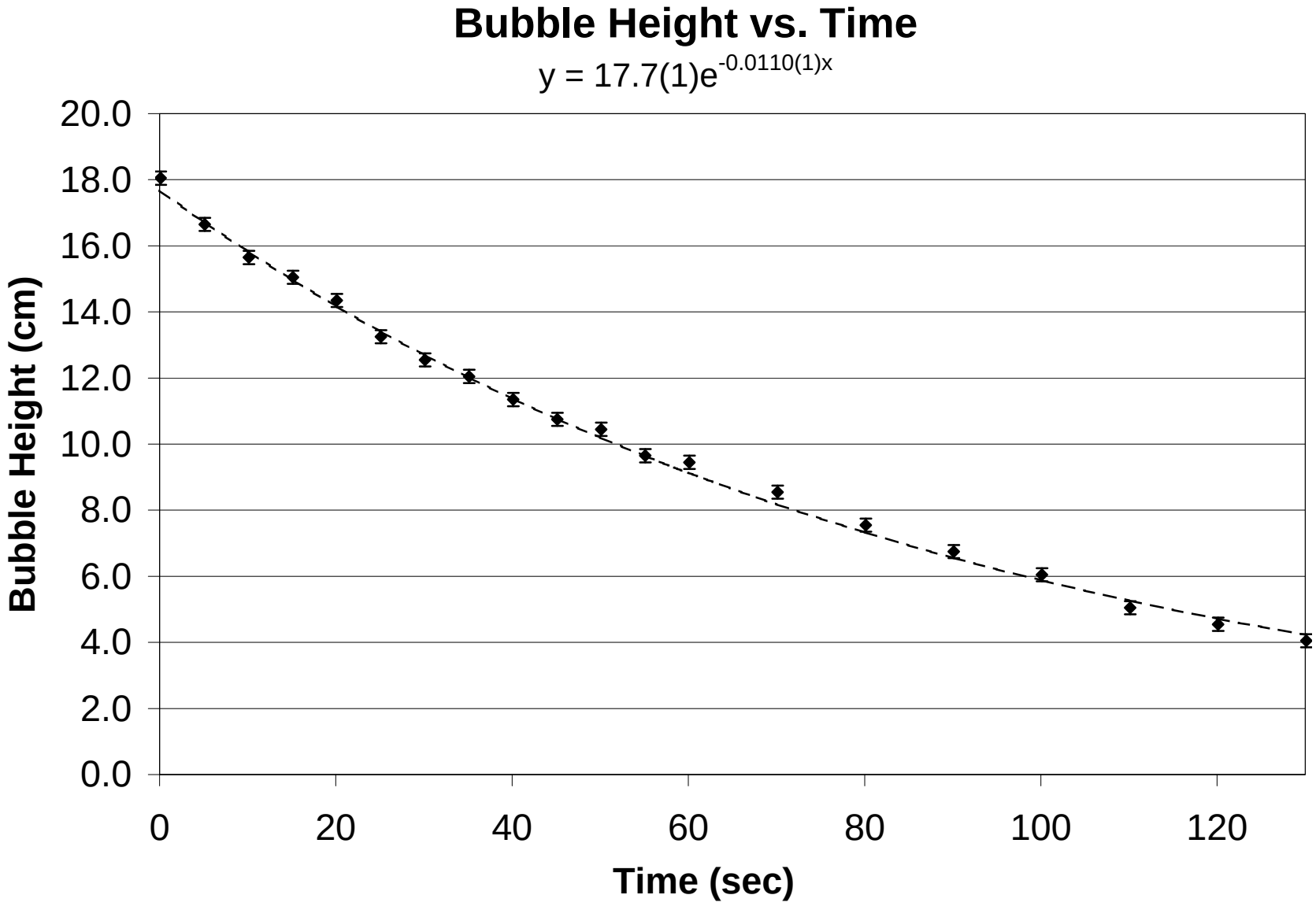
Closing Remarks

It should be stressed again that this is a description of one way to analyze data using a specific program. There are many other ways in which data analysis can be done, and there are many tools and programs one can use in order to do so. The important point is not the specific routine used in order to analyze data, but the fact that data should be analyzed, the meaning of data analysis, what one essentially does by analyzing data, and the implications of data analysis on theory and scientific method.

References

1. Philip R. Bevington and D. Keith Robinson: Data Reduction and Error Analysis for the Physical Sciences, Third Edition. New York: McGraw-Hill, 2003.

2. K. Schalm: "Least Square Fits to an Arbitrary Function using Excel." <http://capa1.physics.sunysb.edu/~senior/cnindex.html>

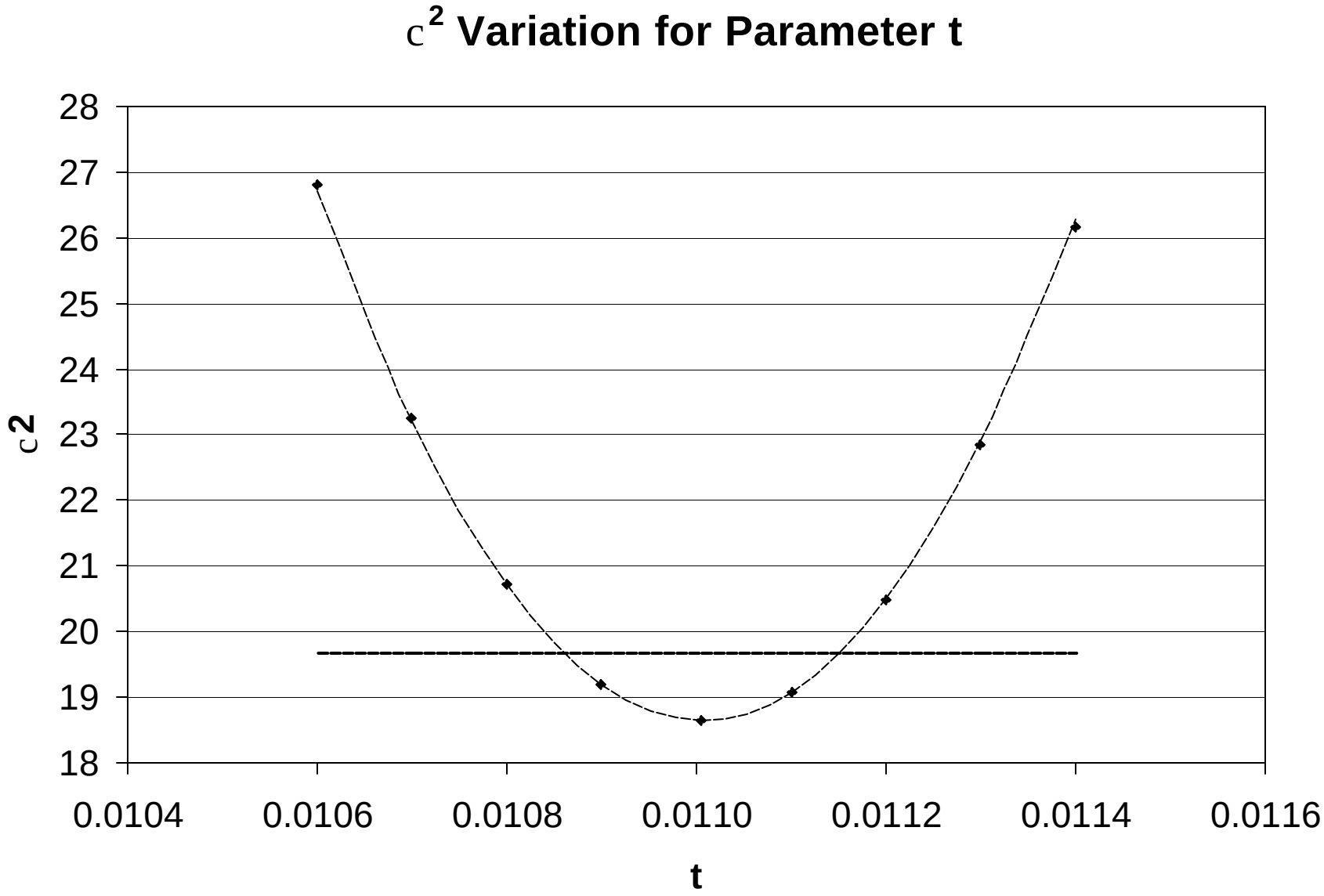

Bubble Height vs. Time
y = 17.7(1)e-0.0110(1)x
Bubble Height (cm)
20.0
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
0
20
40
60
80
100
120
Time (sec)

Exponential Fit : $y = Ne^{tx}$ $y = 17.7(1)e^{-0.0110(1)x}$

N	t	χ^2	Norm χ^2
17.67716	-0.0110053	18.64302	1.035723

time (sec)	height (cm)	σ height (cm)	Fit	Norm Resid
0	18.0	0.2	17.67716	1.614225
5	16.6	0.2	16.73072	-0.653607
10	15.6	0.2	15.83496	-1.174798
15	15.0	0.2	14.98716	0.064215
20	14.3	0.2	14.18475	0.576272
25	13.2	0.2	13.4253	-1.126477
30	12.5	0.2	12.70651	-1.03253
35	12.0	0.2	12.0262	-0.131002
40	11.3	0.2	11.38232	-0.411592
45	10.7	0.2	10.77291	-0.36455
50	10.4	0.2	10.19613	1.019355
55	9.6	0.2	9.650229	-0.251145
60	9.4	0.2	9.133556	1.332218
70	8.5	0.2	8.181718	1.59141
80	7.5	0.2	7.329074	0.854631
90	6.7	0.2	6.565287	0.673567
100	6.0	0.2	5.881096	0.59452
110	5.0	0.2	5.268207	-1.341037
120	4.5	0.2	4.71919	-1.09595
130	4.0	0.2	4.227387	-1.136937

	A	B	C	D	E	F	G
1	Exponential Fit :		$y = Ne^{tx}$				
2							
3		N	t		χ^2	Norm χ^2	
4		17.67717	-0.011005		**=SUMSQ(F8:F27)**	**=E4/(20-2)**	
5							
6		time	height	σ height	Fit	Norm	
7		(sec)	(cm)	(cm)		Resid	
8		0	18.0	0.2	**=$B$4*EXP($C$4*B8)**	**=(C8-E8)/D8**	
9		5	16.6	0.2	16.73073451	-0.653672549	
10		10	15.6	0.2	15.83497211	-1.174860563	
11		15	15.0	0.2	14.98716877	0.064156131	
12		20	14.3	0.2	14.18475677	0.576216157	
13		25	13.2	0.2	13.42530585	-1.12652924	
14		30	12.5	0.2	12.70651588	-1.032579393	
15		35	12.0	0.2	12.02620988	-0.131049378	
16		40	11.3	0.2	11.38232741	-0.41163705	
17		45	10.7	0.2	10.77291837	-0.364591841	
18		50	10.4	0.2	10.19613705	1.019314773	
19		55	9.6	0.2	9.650236556	-0.251182779	
20		60	9.4	0.2	9.133563542	1.332182292	
21		70	8.5	0.2	8.181724372	1.591378139	
22		80	7.5	0.2	7.329079543	0.854602287	
23		90	6.7	0.2	6.565291679	0.673541604	
24		100	6.0	0.2	5.8811007	0.594496498	
25		110	5.0	0.2	5.268211549	-1.341057747	
26		120	4.5	0.2	4.719193624	-1.095968121	
27		130	4.0	0.2	4.227390691	-1.136953456	
28							


c^2 Variation for Parameter t
28
27
26
25
24
23
22
21
20
19
18
c2
0.0104
0.0106
0.0108
0.0110
0.0112
0.0114
0.0116
t

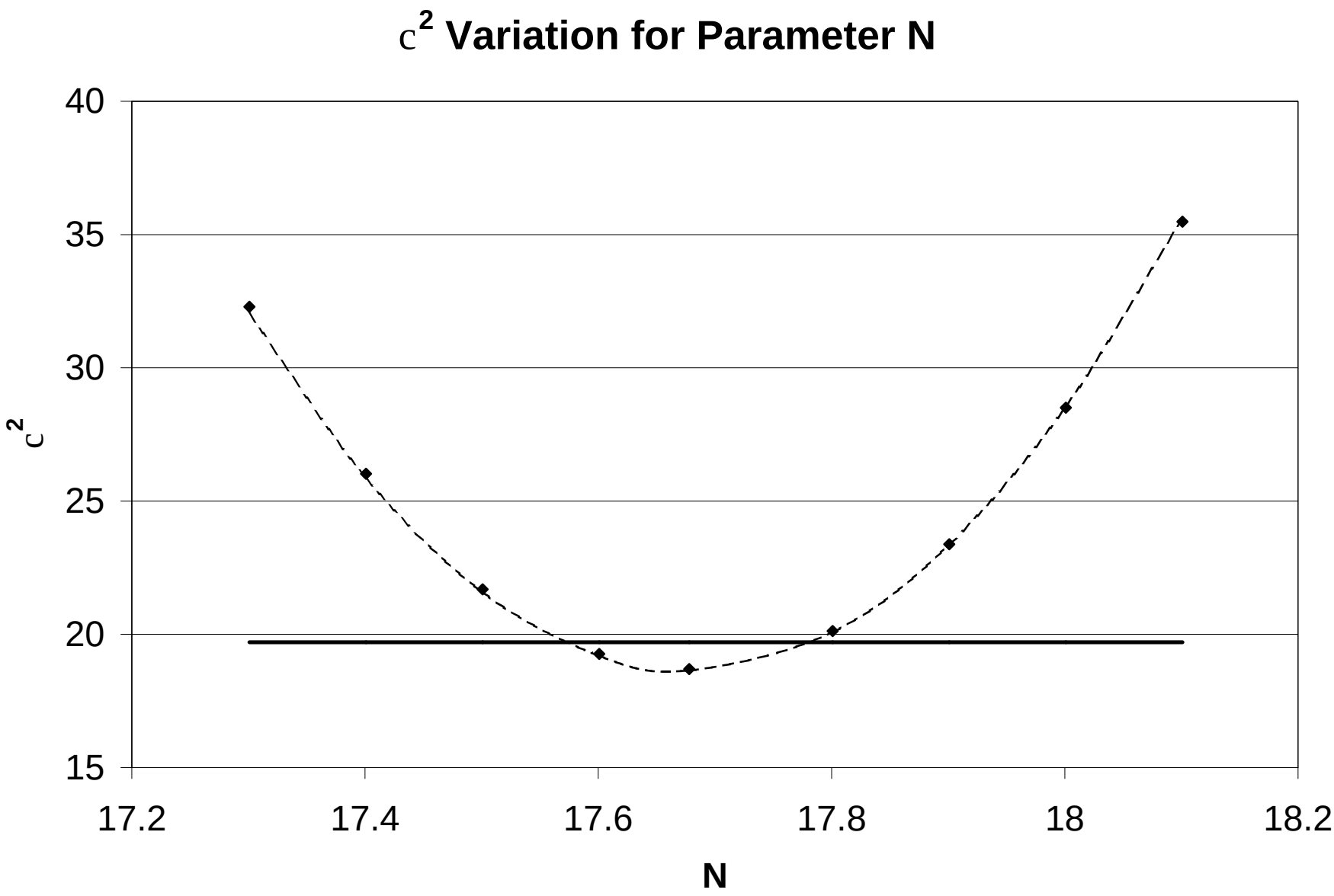
Parabolic Fit : $y = c(x - t)^2 + (\chi^2)$

Values From Exp. Fit

t	χ^2	c	Sum Dev Sq
0.011005	18.64302	49051589.9	0.030660019

varying t	corresp. χ^2	parabolic fit	deviation
0.0106	26.81118	26.70062	0.11055062
0.0107	23.2607	23.21502	0.04567847
0.0108	20.72331	20.71045	0.01285968
0.0109	19.18814	19.18691	0.00123555
0.011005	18.64302	18.64302	0
0.0111	19.08124	19.08292	-0.00168297
0.0112	20.48797	20.50248	-0.01450834
0.0113	22.85387	22.90307	-0.04919726
0.0114	26.16827	26.28468	-0.11641076

	A	B	C	D	E	F
1	Parabolic Fit :	$y = c(x - t)^2 + (\chi^2)$				
2						
3		Values From Exp. Fit				
4		t	χ^2		c	Sum Dev Sq
5		0.011005	18.64302		49051589.9	**=SUMSQ(E9:E17)**
6						
7		varying	corresp.	parabolic	deviation	
8		t	χ^2	fit		
9		0.0106	26.81118	**=$D$5*(B9-$A$5)^2+$B$5**	**=C9-D9**	
10		0.0107	23.2607	23.21502058	0.04567847	
11		0.0108	20.72331	20.71044852	0.01285968	
12		0.0109	19.18814	19.18690826	0.00123555	
13		0.011005	18.64302	18.64302205	0	
14		0.0111	19.08124	19.08292314	-0.00168297	
15		0.0112	20.48797	20.50247827	-0.01450834	
16		0.0113	22.85387	22.9030652	-0.04919726	
17		0.0114	26.16827	26.28468393	-0.11641076	
18						


c^2 Variation for Parameter N
40
35
30
25
20
15
c^2
17.2
17.4
17.6
17.8
18
18.2
N

Parabolic Fit : $y = c(x - N)^2 + (\chi^2)$

Values From Exp. Fit

N	χ^2	c	Sum Dev Sq
17.67716	18.64302	94.53202522	0.040647403

varying N	corresp. χ^2	parabolic fit	deviation
17.3	32.2347	32.08982	0.144880759
17.4	25.96615	25.90449	0.061653856
17.5	21.62815	21.60981	0.018339454
17.6	19.20785	19.20576	0.002089304
17.67716	18.64302	18.64302	0
17.8	20.06952	20.06959	-7.29767E-05
17.9	23.32632	23.33747	-0.011148258
18.0	28.4505	28.49599	-0.045488784
18.1	35.42972	35.54514	-0.115419027

	A	B	C	D	E	F
1	Parabolic Fit :	$y = c(x - N)^2 + (\chi^2)$				
2						
3		Values From Exp. Fit				
4		N	χ^2		c	Sum Dev Sq
5		17.67716	18.64302		94.5320252	**=SUMSQ(F9:F17)**
6						
7		varying	corresp.	parabolic	deviation	
8		N	χ^2	fit		
9		17.3	32.2347	**=$F$5*(C9-$C$5)^2+$D$5**	**=D9-E9**	
10		17.4	25.96615	25.90449305	0.06165386	
11		17.5	21.62815	21.60980735	0.01833945	
12		17.6	19.20785	19.20576216	0.0020893	
13		17.67716	18.64302	18.64302205	0	
14		17.8	20.06952	20.06959327	-7.298E-05	
15		17.9	23.32632	23.33746959	-0.0111483	
16		18.0	28.4505	28.49598641	-0.0454888	
17		18.1	35.42972	35.54514374	-0.115419	
18						